NYCU-NLP at EXIST 2024: Leveraging Transformers with Diverse Annotations for Sexism Identification in Social Networks

Notebook for the EXIST Lab at CLEF 2024

Yi-Zeng Fang[1], Lung-Hao Lee[2,*] and Juinn-Dar Huang[1]

[1]Institute of Electronics, National Yang Ming Chiao Tung University, Taiwan
[2]Institute of Artificial Intelligence Innovation, National Yang Ming Chiao Tung University, Taiwan

Abstract

This paper presents a robust methodology for identifying sexism in social media texts as part of the EXIST 2024 challenge. First, we incorporate extensive data preprocessing techniques, including removing redundant elements, standardizing text formats, increasing data diversity by the back-translation, and augmenting texts using the AEDA approach. We then integrate annotator demographics such as gender, age, and ethnicity into our selected transformer-based language models. The rounding technique is used to handle non-continuous annotation values to maintain precise probability distributions. We empirically optimize shared layers across tasks based on the hard parameter-sharing techniques to improve generalization and computational efficiency. Rigorous evaluations were conducted using five-fold cross-validation to ensure the reliability of the findings. Finally, our system was respectively ranked first out of 40, 35, and 33 submissions for Tasks 1, 2 and 3 in the Soft-Soft category setting. In addition, in the Hard-Hard category setting, our system was ranked the first out of 70 submissions for Task 1; second out of 46 submissions for Task 2; and third out of 34 submissions for Task 3. This paper reports our findings in classifying sexism within social media textual content, offering substantial insights for the EXIST 2024 challenge.

Keywords

Sexism Identification, Pre-trained Language Models, Text Classification, Transformers

1. Introduction

Social media platforms like Twitter, Instagram, and Facebook have become integral to modern communication and information sharing. However, these platforms also facilitate the spread of discriminatory and prejudiced content, such as sexism. Sexism is a form of discrimination based on gender that undermines the dignity and rights of women and marginalized groups through insults, stereotypes, jokes, threats, and harassment. Identifying and filtering objectionable web content is crucial for fostering a respectful and inclusive online environment [1].

The EXIST (sEXism Identification in Social neTworks) is a series of shared tasks to capture instances of sexism, ranging from explicit misogyny to other subtle expressions that involve implicit sexist behaviors [2, 3, 4]. The EXIST 2024 [5, 6] challenge contains three traditional tasks for classifying sexist textual messages. Task 1 (Sexism Identification in Tweets): This is a binary task used to decide whether a tweet contains sexist expressions or behaviors. Task 2 (Source Intention in Tweets): This is a multi-class task used to classify tweets identified as sexist in Task 1 into three categories based on the author's intention, including 1) Direct: the tweet itself is a sexist message; 2) Reported: the tweet reports or describes a sexist event or situation; and 3) Judgmental: the tweet condemns sexist situations or behaviors. Task 3 (Sexism Categorization in Tweets): This is a multi-label task used to further categorize tweets identified as sexist into defined types, including 1) Ideological-Inequality: discrediting feminism or presenting men as victims of gender inequality; 2) Stereotyping-Dominance: promoting traditional gender roles or suggesting male superiority; 3) Objectification: treating women as objects, often focusing on physical

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
*Corresponding author.
joycefang1213.ee11@nycu.edu.tw (Y. Fang); lhlee@nycu.edu.tw (L. Lee); jdhuang@nycu.edu.tw (J. Huang)

appearance or traditional gender roles; 4) Sexual-Violence: including sexual suggestions, harassment, or assault; and 5) Misogyny-Non-Sexual-Violence: expressing hatred or non-sexual violence towards women. The EXIST-2024 datasets contain tweets in English and Spanish annotated with sexist remarks. Similar to the 2023 edition, this edition also embraces the Learning With Disagreement paradigm for dataset development and system evaluations. Therefore, developed systems can learn from conflicting or diverse annotations, allowing for a fairer learning process by considering the perspectives, biases, or interpretations of multiple annotators. Given the success of transformer models in various NLP tasks, our approach explores the use of transformer-based language models to identify and classify tweets for sexism detection.

This paper describes the NYCU-NLP system for the EXIST 2024 challenge. We use extensive data preprocessing techniques, including removing irrelevant elements, standardizing text formats, back-translation via the Google Translator API, and implementing the AEDA [7] method for text augmentation. We also adapt the Round to Closed Value method [8] to handle non-continuous annotation values. The main system architecture is the transformer-based language model. We integrate annotator information, such as gender, age, and ethnicity, to create a unified vector representation for each tweet. This integration enriches the model's contextual understanding and improves its ability to identify sexist content. We further incorporate Hard Parameter Sharing [9] to optimize shared layers across tasks to enhance generalization and computational efficiency. We rigorously evaluate our model performance through 5-fold cross-validation to ensure reliability and minimize over-fitting. Finally, our model obtained outstanding performance in the EXIST 2024 challenge, ranking respectively first out of 40, 35, and 33 submissions for Tasks 1, 2, and 3 with the Soft-Soft category setting. In addition, in the Hard-Hard category configuration, our system was ranked first out of 70 submissions for Task 1, second out of 46 submissions for Task 2, and third out of 34 submissions for Task 3. Our findings reflect ongoing efforts to detect and categorize sexism in social media, offering valuable insights for the EXIST 2024 challenge and beyond.

The rest of this paper is organized as follows. Section 2 investigates related studies for sexism identification. Section 3 describes the NYCU-NLP system for the EXIST-2024 tasks. Section 4 presents results and performance comparisons. Conclusions are finally drawn in Section 5.

2. Related Work

The automated detection of sexism on digital platforms has become increasingly important due to its prevalence and the sheer volume of content needing review, necessitating the development of systems to quicking and effectively identify and counteract such content. Researchers have explored various methods, initially focusing on rule-based systems but now predominantly using machine learning techniques, particularly pre-trained transformer models like BERT and its derivatives [10, 11, 12, 13]. These advanced models now outperform traditional methods in capturing sexist language's contextual and semantic nuances.

Despite these advancements, sexism detection remains challenging due to the subjective and culturally variable nature of sexist behavior. Initiatives such as the EXIST 2023 [4] and SemEval 2023 challenges [14] have emphasized the need for detailed classification systems, introducing taxonomies that categorize sexism into distinct types, including ideological sexism, stereotyping, and misogyny. These taxonomies aim to enhance the explainability and comprehensiveness of sexism detection systems.

Bias in detection models is another critical issue. Models can perpetuate biases in their training data, leading to skewed results. Recent studies have addressed this by incorporating perspectivism and analyzing annotator agreement, which can improve the fairness and accuracy of these systems [15, 16, 17]. This consideration is particularly important in multilingual contexts, where expressions of sexism can vary widely.

While machine learning and deep learning models have significantly advanced sexism detection, challenges remain in addressing bias, subjectivity, and the diverse forms of sexism across cultures and media types. Further research is needed, leveraging multimodal analysis and incorporating nuanced,

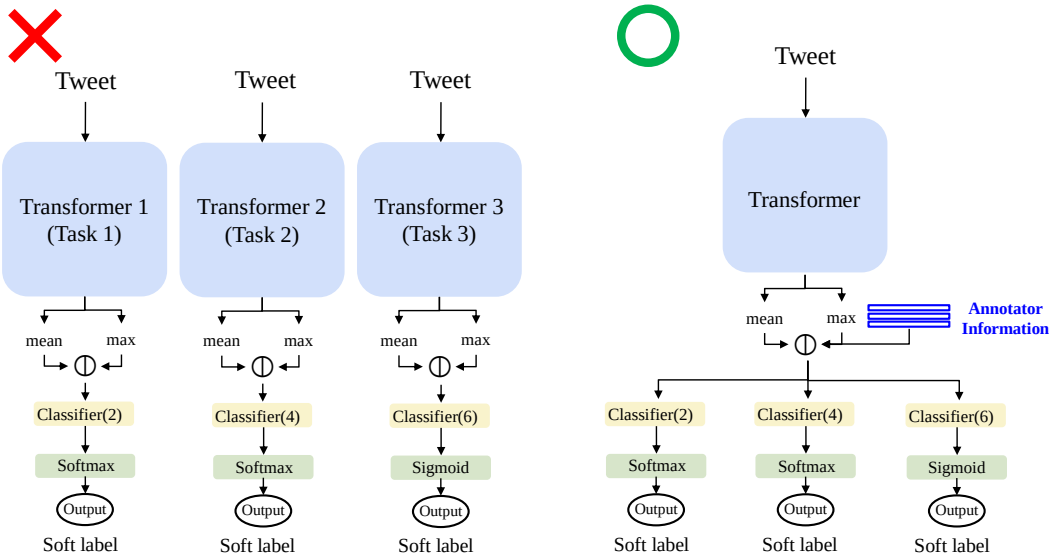
context-aware approaches to develop more robust detection systems.

3. The NYCU-NLP System

We use Hard Parameter Sharing [9] to efficiently train our transformer-based model on tasks that exhibit an inherent hierarchical relationship, specifically Tasks 1 through 3. This technique involves sharing hidden layers across all tasks, enabling the network to learn a common representation that leverages the shared features of these related tasks. Given the sequential nature of our tasks, where each task builds upon the preceding one, training them in isolation would be sub-optimal and could result in redundant or conflicting representations. We show the parameter sharing architecture in Fig. 1 (b).

Hard Parameter Sharing ensures that the foundational knowledge acquired from Task 1 is effectively utilized and refined in subsequent tasks, thereby enhancing the model's overall performance and generalization. This approach mitigates the risk of over-fitting through the regularizing effect of shared parameters and improves computational efficiency by reducing the number of required parameters compared to training separate models for each task. Consequently, Hard Parameter Sharing is a suitable and effective method for our multi-task learning scenario.

To prepare the data for analysis during the pre-processing phase, we first removed usernames, URLs, percentages, time, dates, hashtags, and emojis, as these elements were unlikely to influence the annotators' judgments (see Fig. 2). Subsequently, all characters are converted to lowercase to ensure uniformity and reduce the complexity of the text data. We also translated the text from English to Spanish and then back to English via the Google Translator API, effectively doubling the amount of data and introducing subtle variations that can improve the robustness of our models, as shown in Fig. 3.



(a) wo/ share (b) w/ share

Figure 1: Schematic representation of our model architecture using Hard Parameter Sharing across three tasks. This diagram illustrates the shared transformer layers and specific classifiers for each task, highlighting the integration of annotator information and the utilization of softmax and sigmoid outputs. The comparative setup (a) without sharing versus (b) with sharing underscores the benefits of shared parameters in reducing redundancy and enhancing task interdependencies.

3.1. Data Augmentation

We use the AEDA technique [7] to augment the text data by randomly segmenting sentences and inserting punctuation marks from a predefined set ".", ";", "?", ":", "!", ",". AEDA offers advantages

ID	Lang.	Original Sentence	Transform Sentence
100199	ES	¡Gran oportunidad de exposición para sus juegos en @Steam! @wingsfundme está aceptando propuestas de videojuegos que ya estén presentes en #Steam para figurar en el evento de #WomensDay 🙋‍♀️ Pueden ver todos los requisitos para participar en el formulario del tweet 😘 https://t.co/3W6PtUTDdR	¡gran oportunidad de exposición para sus juegos en ! está aceptando propuestas de videojuegos que ya estén presentes en para figurar en el evento de pueden ver todos los requisitos para participar en el formulario del tweet
200176	EN	Feel #blessed that I have raised a caring & loving 13 yo who is our Next Gen Feminist & Ally. I was crying of joy inside when I got this text. Not only we must #BreakTheBias for women, we need to do it for our children. 🗼👄🎈🍿 @GlobalFundWomen @UN_Women @womensday @WomeninID https://t.co/UJvvloR0iP	feel that i have raised a caring & loving 13 yo who is our next gen feminist & ally. i was crying of joy inside when i got this text. not only we must for women, we need to do it for our children.

Figure 2: Data cleaning process, including the removal of usernames, URLs, emojis, hashtags, and other non-essential elements, followed by conversion of all texts to lowercase. This figure showcases the transformation of example tweets, highlighting the streamlined and standardized text that forms the basis for further analysis.

ID	Lang. / Original Sentence	Lang. / New Sentence
100199	ES	EN
	¡gran oportunidad de exposición para sus juegos en ! está aceptando propuestas de videojuegos que ya estén presentes en para figurar en el evento de pueden ver todos los requisitos para participar en el formulario del tweet	Great exposure opportunity for your games on ! is accepting proposals for video games that are already present in to appear in the event. You can see all the requirements to participate in the tweet form
200176	EN	ES
	feel that i have raised a caring & loving 13 yo who is our next gen feminist & ally. i was crying of joy inside when i got this text. not only we must for women, we need to do it for our children.	Siento que he criado una familia cariñosa y amable. cariñosa de 13 años, que es nuestra feminista y campesina de próxima generación. aliado. Estaba llorando de alegría por dentro cuando recibí este mensaje. No sólo debemos hacerlo por las mujeres, sino que debemos hacerlo por nuestros hijos.

Figure 3: Data augmentation process via back-translation using the Google Translator API. This figure presents original English sentences translated to Spanish and then back to English, illustrating the subtle variations introduced to enhance the dataset's diversity and robustness for model training.

including its simplicity and effectiveness in generating diverse textual variations without significantly altering the semantic content. Unlike traditional text augmentation techniques [18] such as synonym replacement, random insertion, random swap, and random deletion, which may cause unintended biases and distortions, the AEDA technique maintains the integrity of the original data, ensuring that the augmented dataset remains representative of the original distribution, thereby enhancing the robustness and generalizability of our model.

3.2. Incorporating Annotator Information

Each tweet was annotated by up to six annotators, and their demographic information was stored in the EXIST 2024 datasets [5, 6]. We converted each annotator's gender, age, and ethnicity information into one-hot encoded vectors, transforming categorical variables into a binary vector representation suitable for input into our neural network model. Each one-hot encoded vector [19] is passed through an embedding layer to obtain a dense 16-dimensional representation. The embedding layer is trained to map similar categories closer in the vector space, capturing the underlying relationships between different annotator attributes. For each tweet, we average the sum of the 16-dimensional embedding vectors [20] of the six annotators, resulting in a single 16-dimensional vector representing the combined annotator information.

3.3. Round to Closed Value

We use Round to Closed Value [8] to ensure the output probability is close to the real value that matches the number of annotators. The method was applied uniformly to Task 1 and Task 2, as they both involve a mono-label classification where the sum of probabilities should be 1. We first generate all possible probability combinations for the given labels. For example, [1/6, 5/6] is a valid combination for Task 1 with 2 categories. We then calculate the cosine similarity between these valid combinations and the model's predicted probabilities. The combination is most similar to the prediction chosen as the adjusted value.

We modify the Round to Closed Value [8] approach for Task 3, which is a multi-label classification task where the sum of probabilities exceeds 1. We use the minimum of absolute differences to find the closest value for adjustment. The total adjusted probability might be below 1. We then select the next closest category and adjust its probability accordingly, ensuring the sum of adjusted probabilities is at least 1. This step ensures that the cumulative probability is valid and meaningful.

4. Evaluation

4.1. Datasets

The EXIST 2024 Tweets Datasets [5, 6] aim to facilitate the identification and analysis of sexism in social media content. Table 1 shows the datasets comprising over 10,000 annotated tweets in both English and Spanish with a balanced distribution. Each tweet in the dataset is represented as a JSON object containing the following attributes: 1) id_EXIST: unique identifier for the tweet; 2) lang: language of the tweet text ("en" for English or "es" for Spanish); 3) tweet: text content of the tweet; 4) number_annotators: number of annotators who labeled the tweet; 5) annotators: unique identifiers for each annotator; 6) gender_annotators: gender of the annotators (values: "F" for female and "M" for male); 7) age_annotators: age group of the annotators (values: "18-22", "23-45", "46+"); 8) ethnicity_annotators: ethnicity of the annotators (e.g., "Black or African American", "Hispano or Latino", etc.); 9) study_level_annotators: educational level of the annotators (e.g., "high school degree or equivalent", "bachelor's degree", etc.); 10) country_annotators: country where the annotators reside; 11) labels_task1: one label indicates whether the tweet contains sexist content (values: "yes" or "no"); 12) labels_task2: one label categorizes the intention behind the sexist tweet (values: "direct", "reported", "judgemental", "-", "unknown"); 13) labels_task3: one label indicates the type(s) of sexism present in the tweet, if any (e.g., "ideological-inequality", "stereotyping-dominance", etc.)

The dataset is annotated by a diverse group of individuals in terms of gender, age, ethnicity, education level, and country of residence, enhancing the robustness and fairness of the annotations. This diversity helps ensure the dataset captures various perspectives and reduces potential annotation biases.

4.2. Settings

We use the five-fold cross-validation technique to evaluate the model performance during development. This method involves partitioning the combined datasets, including training and development data, into five folds of equal size. During each iteration, one fold is designated as the validation set, while

Table 1
Distribution of the dataset across different phases of model training, segmented by language. The table details the number of instances in the training, development, and test sets for both Spanish and English, highlighting the balanced allocation to support effective model generalization across both languages.

Language	Training	Development	Test
Spanish	3660	549	1098
English	3260	489	978
Total	6920	1038	2076

the remaining four folds are used for model training. This process is repeated five times, ensuring that each fold is used exactly once as the validation set. We derive a robust estimate of the model's generalization capability by averaging the performance metrics obtained from each iteration. The use of five-fold cross-validation not only maximizes the utility of our datasets but also provides a reliable means of assessing the model's performance, reducing the potential for over-fitting and ensuring that the evaluation is not biased to any single train-test split.

DeBERTaV3-large [12] and XLM-RoBERTa-large [13] were used as main transformer models across three tasks. The hyperparameters were empirically configured as follows. We used the AdamW optimizer [21] for training, with a learning rate of 1e-5 and a dropout rate of 0.1. The training process spanned 30 epochs, with a maximum sequence length of 128 tokens. A batch size of 20 was used to optimize performance across the tasks. The evaluation framework was implemented on a single NVIDIA Tesla V100 GPU with 32GB of memory.

Our training primarily focused on the Soft-Soft category setting only, so the Hard-Hard category setting was not specifically trained on. Tasks 1 and 2 used a direct conversion with maximum values, whereas Task 3 used a conversion threshold of 0.16666. The ICM-Soft values [22] for each task were measured using the PyEvALL Evaluation Library, with results averaged over five-fold cross-validation.

4.3. Results

Tables 2 and 3 show the respective five-fold cross-validation results using DeBERTaV3-large [12] and XLM-RoBERTa-large [13]. Various configurations were tested, including data augmentation (denoted as DA), annotator information (AI), rounding to closed values (RC) [8], and translation from Spanish to English (Tr.). The configurations denoted as V1, V2, and V3 represent the final versions submitted for official evaluation.

The DeBERTaV3-large model consistently surpassed the XLM-RoBERTa-large model across all tasks and configurations. The performance disparity was particularly pronounced in the final versions (V2 for DeBERTaV3-large and V3 for XLM-RoBERTa-large). Specifically, DeBERTaV3-large demonstrates superior results in Task 1 (1.0084 vs. 0.9370) and achieves lower negative values in Task 2 (-0.5208 vs. -0.9049) and Task 3 (-1.8042 vs. -2.3777).

These findings reveal the effectiveness of integrating data augmentation, annotator information, the rounding to closed values technique, and the Spanish-to-English translation. Overall, the DeBERTaV3-large model in its version 2 (V2) emerged as the most effective model for the tasks assessed in this study.

Table 2
Performance results of the DeBERTaV3-large model across three tasks.

DeBERTaV3-large	Task 1 ↑	Task 2 ↑	Task 3 ↑
baseline	0.7849	-1.2073	-3.2058
+ DA	0.8287	-0.9153	-2.9269
+ DA + AI	0.9410	-0.7256	-2.5230
+ DA + AI + RC (V1)	0.9862	-0.5597	-1.9450
+ DA + AI + RC + Tr. (V2)	**1.0084**	**-0.5208**	**-1.8042**

Table 3
Performance evaluation of the XLM-RoBERTa-large model across three tasks.

XLM-RoBERTa-large	Task 1 ↑	Task 2 ↑	Task 3 ↑
baseline	0.7605	-1.5976	-3.5432
+ DA	0.8063	-1.3867	-3.3798
+ DA + AI	0.9072	-0.9970	-2.8341
+ DA + AI + RC (V3)	**0.9370**	**-0.9049**	**-2.3777**
+ DA + AI + RC + Tr.	0.9005	-0.9251	-2.5723

4.4. Rankings

Tables 4, 5 and 6 respectively show our final submissions with the Soft-Soft and Hard-Hard category settings on the test set. In the Soft-Soft category setting, our model was respectively ranked first for Tasks 1, 2, and 3 out of 40, 35, and 33 submissions. In the Hard-Hard category setting, our system was ranked first out of 70 submissions for Task 1, second out of 46 submissions for Task 2, and third out of 34 submissions for Task 3. In summary, our findings reflect ongoing efforts to detect and categorize sexism in social media.

Table 4
Final results on the test set for Task 1

Lang	Version	Soft Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Hard Rank	ICM-Hard	ICM-Hard Norm	Macro F1
All	1	1	**1.0944**	**0.6755**	**0.9088**	1	**0.5973**	**0.8002**	**0.7944**
	2	2	1.0866	0.6742	0.8826	9	0.5619	0.7824	0.7785
	3	3	1.0810	0.6733	0.9831	8	0.5749	0.7889	0.7813
ES	1	1	**1.1434**	**0.6834**	**0.8681**	1	**0.6215**	**0.8108**	**0.8238**
	2	3	1.1251	0.6804	0.8751	8	0.5805	0.7903	0.8077
	3	2	1.1358	0.6822	0.9229	5	0.5995	0.7998	0.8075
EN	1	2	1.0024	0.6609	0.9545	5	**0.5564**	**0.7839**	**0.7557**
	2	1	**1.0158**	**0.6631**	**0.8911**	13	0.5298	0.7704	0.7410
	3	3	0.9841	0.6580	1.0506	11	0.5362	0.7736	0.7477

Table 5
Final results on the test set for Task 2

Lang	Version	Soft Rank	ICM-Soft	ICM-Soft Norm	Cross Entropy	Hard Rank	ICM-Hard	ICM-Hard Norm	Macro F1
All	1	2	-0.4059	0.4673	1.8549	3	0.3383	0.6100	0.5353
	2	1	**-0.2543**	**0.4795**	**1.8344**	4	0.3073	0.5999	0.5273
	3	3	-0.5226	0.4579	1.9206	2	**0.3522**	**0.6145**	**0.5410**
ES	1	2	-0.2633	0.4789	1.8228	1	**0.4457**	**0.6392**	**0.5757**
	2	1	**-0.0756**	**0.4939**	**1.8197**	4	0.4098	0.6280	0.5723
	3	3	-0.3308	0.4735	1.8540	3	0.4113	0.6285	0.5697
EN	1	2	-0.6464	0.4472	1.8909	4	0.1881	0.5651	0.4729
	2	1	**-0.5041**	**0.4588**	**1.8509**	6	0.1692	0.5585	0.4625
	3	3	-0.8235	0.4327	1.9954	2	**0.2672**	**0.5925**	**0.4991**

Table 6
Final results on the test set for Task 3

Lang	Version	Soft Rank	ICM-Soft	ICM-Soft Norm	Hard Rank	ICM-Hard	ICM-Hard Norm	Macro F1
All	1	1	**-1.1762**	**0.4379**	4	0.2364	0.5549	0.6066
	2	2	-1.2169	0.4357	5	0.1725	0.5401	0.5933
	3	3	-1.4555	0.4231	3	**0.3069**	**0.5713**	**0.6130**
ES	1	1	**-1.1280**	**0.4413**	4	0.2986	0.5667	0.6206
	2	2	-1.1584	0.4397	5	0.1653	0.5369	0.5968
	3	3	-1.2881	0.4330	3	**0.3138**	**0.5701**	**0.6228**
EN	1	1	**-1.2583**	**0.4311**	5	0.1448	0.5355	0.5855
	2	2	-1.2802	0.4299	4	0.1680	0.5412	0.5874
	3	3	-1.7322	0.4051	1	**0.2820**	**0.5691**	**0.5989**

5. Conclusions

This study describes the NYCU-NLP submission for the EXIST-2024 Tasks 1, 2 and 3, including system design, implementation and evaluation. We remove superfluous elements, standardize the text formats, increase data diversity by the back-translation, and augment texts using the AEDA approach. We then integrate annotator demographics such as gender, age, and ethnicity into our selected transformer-based language models. Our model architecture based on the Hard Parameter Sharing technique optimizes computational efficiency and improves performance by leveraging shared features across related tasks. The results of the EXIST 2024 challenge demonstrate that our methodology significantly improves the detection and categorization of sexism in social media. Our approach yielded excellent performance, underscoring the effectiveness of the advanced techniques and strategies implemented.

Acknowledgments

This work was partially supported by the Ministry of Science and Technology, Taiwan, under grant MOST-111-2218-E-A49-022, and the National Science and Technology Council, Taiwan, under grant NSTC 111-2628-E-A49-029-MY3. We also thank the National Center for High-performance Computing and Taiwan Computing for supporting computing resources.

References

[1] L.-H. Lee, Y.-C. Juan, W.-L. Tseng, H.-H. Chen, Y.-H. Tseng, Mining browsing behaviors for objectionable content filtering, Journal of the Association for Information Science and Technology 66 (2015) 930–942.

[2] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.

[3] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.

[4] L. Plaza, J. Carrillo-de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023: sexism identification in social networks, in: European Conference on Information Retrieval, Springer, 2023, pp. 593–599.

[5] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[6] L. Plaza, J. Carrillo-de-Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024 – Learning with Disagreement for Sexism Identification and Characterization in Social Networks and Memes (Extended Overview), in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum, 2024.

[7] A. Karimi, L. Rossi, A. Prati, Aeda: an easier data augmentation technique for text classification, arXiv preprint arXiv:2108.13230 (2021).

[8] A. F. M. de Paula, G. Rizzi, E. Fersini, D. Spina, Ai-upv at exist 2023–sexism characterization using large language models under the learning with disagreements regime, arXiv preprint arXiv:2307.03385 (2023).

[9] S. Ruder, An overview of multi-task learning in deep neural networks, arXiv preprint arXiv:1706.05098 (2017).

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[11] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942 (2019).
[12] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).
[14] A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. T. Madabushi, R. Kumar, E. Sartori, Proceedings of the 17th international workshop on semantic evaluation (semeval-2023), in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023.
[15] S. Raza, O. Bamgbose, V. Chatrath, S. Ghuge, Y. Sidyakin, A. Y. Muaad, Unlocking bias detection: Leveraging transformer-based models for content analysis, arXiv preprint arXiv:2310.00347 (2023).
[16] A. Radwan, L. Zaafarani, J. Abudawood, F. AlZahrani, F. Fourat, Addressing bias through ensemble learning and regularized fine-tuning, arXiv preprint arXiv:2402.00910 (2024).
[17] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira, et al., Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods, Big data and cognitive computing 7 (2023) 15.
[18] J. Wei, K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks, arXiv preprint arXiv:1901.11196 (2019).
[19] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities., Proceedings of the national academy of sciences 79 (1982) 2554–2558.
[20] F. Almeida, G. Xexéo, Word embeddings: A survey, arXiv preprint arXiv:1901.09069 (2019).
[21] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).
[22] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 5809–5819.